This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549

Wireless Vision-Centered Semantic Communication for Smart City Environment: Pretrained Network and Quantization

Yan Gong, Zheng Chu, *Member, IEEE*, Zhengyu Zhu, *Senior Member, IEEE*, Pei Xiao, *Senior Member, IEEE*, Ming Zeng, *Member, IEEE*, Yi Wang, Hari Mohan Pandey, and Jingrui Hou

***Abstract*—This paper introduces a Vision-Centered Semantic Communication (VCSC) system tailored for efficient image transmission in smart city environments, where bandwidth is limited and channels are subject to severe noise. Unlike conventional text-centered or classical compression approaches, VCSC leverages a pretrained latent encoder–decoder network to extract compact, semantically rich representations directly from images. An innovative attention-based quantization strategy is employed to selectively allocate higher precision to critical regions, thereby reducing the overall bit rate while preserving essential semantic details. The quantized latent codes are robustly transmitted over wireless channels modeled with additive white Gaussian noise and Rayleigh fading. An end-to-end training framework minimizes both reconstruction and perceptual losses, ensuring high-fidelity image recovery even under adverse conditions. Extensive simulations demonstrate that VCSC outperforms traditional methods in preserving fine-grained details and semantic integrity, offering a promising solution for real-time surveillance, transportation, and infrastructure monitoring in smart cities.**

***Index Terms*—Semantic communication, smart city, image transmission, latent code, and quantization.**

The work of Zheng Chu was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. ZCLMS26F0101, and in part by Ningbo Natural Science Foundation under Grant No. 2024J233. The work of Yan Gong and Hari Mohan Pandey was supported in part by the computational resources provided by the School of Computing and Engineering, Bournemouth University. This work of Zhengyu Zhu was supported in part by the National Natural Science Foundation of China under Grant 62571495, 62571182, in part by State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) under Grant SKLNST-2025-1-07. The work of Pei Xiao was supported in part by the U.K. Engineering and Physical Sciences Research Council under Grant EP/X013162/1. The work of Yi Wang was supported in part by the Natural Science Foundation of Henan (No. 252300421516). The work of Jingrui Hou was supported in part by the China Postdoctoral Science Foundation (No. 2025M773205) and the Postdoctoral Fellowship Program of CPSF (No. GZC20252319).

Corresponding author: Zheng Chu.

Yan Gong and Hari Pandey are with the School of Computing and Engineering, Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom. (Email: ygong@bournemouth.ac.uk,hpandey@bournemouth.ac.uk)

Zheng Chu is with the Department of Electrical and Electronic Engineering, University of Nottingham Ningbo China, Ningbo 315100, China. (Email: andrew.chuzheng7@gmail.com)

Zhengyu Zhu is with the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, 450001, China, and is also with State key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China. (Email: zhuzhengyu6@gmail.com)

Pei Xiao is with the 5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, Guildford GU2 7XH, UK (Email: p.xiao@surrey.ac.uk).

Ming Zeng is with the Department of Electrical and Computer Engineering, Laval University, Quebec City G1V 0A6, Canada (email: ming.zeng@gel.ulaval.ca)

Yi Wang is with the School of Electronics and Information, Zhengzhou University of Aeronautics, Zhengzhou 450046, China. (Email: yiwang@zua.edu.cn)

Jingrui Hou is with the School of Information Management, Wuhan University, Wuhan 430072, China. (Email: houjingrui@whu.edu.cn)

I. Introduction

The widespread deployment of high-resolution cameras in smart city applications has led to an exponential increase in visual data transmission demands [1]. Real-time video surveillance, intelligent transportation systems, and automated infrastructure monitoring require efficient image communication over bandwidth-limited and noise-prone wireless networks. These camera networks act as a critical component of urban sensing, continuously capturing visual information to support city-scale perception and decision-making. Conventional image compression methods, such as JPEG [2], struggle to maintain acceptable visual quality under extreme noise conditions or stringent bandwidth constraints [3]. Meanwhile, text-centered semantic communication (TCSC) approaches, which convert images into textual descriptions before reconstruction, often discard critical low-level details, making them unsuitable for tasks requiring precise spatial and texture information [4].

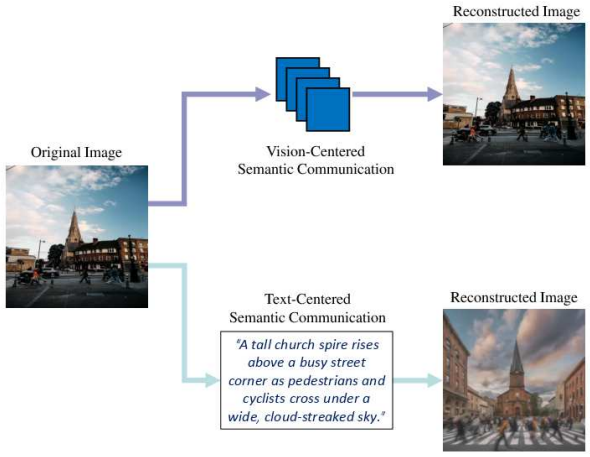


Fig. 1: Comparison of text-centered and vision-centered semantic communication. Vision-centered semantic communication transmits image features to preserve spatial details, while text-centered semantic communication transmits text descriptions, which may discard fine-grained visual details.

To address these challenges, this paper proposes a vision-centered semantic communication (VCSC) system in smart

city environments, as illustrated in Fig. 1. This approach represents a promising paradigm, prioritizing the transmission of essential visual semantics while preserving spatial fidelity. Fig. 1 illustrates the fundamental difference between TCSC and VCSC schemes. Specifically, in the TCSC scheme, an image is first transformed into a textual description, which is then transmitted and used to reconstruct an image. However, this process leads to significant information loss, often resulting in structurally inaccurate and visually distorted outputs. In contrast, the VCSC system directly encodes images into compact latent code, allowing for faithful reconstructions while ensuring robustness against channel impairments. Instead of relying on text-centered descriptions, VCSC leverages a pretrained latent encoder–decoder model to transform images into compact, semantically meaningful representations. These representations are then quantized and transmitted over wireless channels modeled with additive white Gaussian noise (AWGN) and Rayleigh fading. At the receiver, a decoder reconstructs the image by minimizing perceptual loss, thereby ensuring high-quality reconstructions even under challenging transmission conditions.

From an application perspective, VCSC is particularly attractive for bandwidth-limited edge-camera deployments in real-world domains. Typical use cases include city-wide traffic monitoring (e.g., roadside cameras at intersections) and public-safety surveillance, where uplink capacity is constrained but timely and faithful visual reconstruction is required for situational awareness. VCSC can also support smart-city infrastructure monitoring (e.g., bridges, tunnels, and railways) using UAVs or mobile robots over intermittent wireless links, enabling more robust image recovery than conventional codecs under harsh channel conditions.

The main contribution of this paper is summarized as follows:

- **VCSC Framework**: A novel vision-centered semantic transmission framework is proposed, which encodes images into a semantically rich latent space, preserving both structural details and semantic integrity under noisy conditions.
- **Attention-Based Adaptive Quantization**: Unlike uniform quantization, an adaptive quantization strategy dynamically allocates bit precision to semantically significant regions, enhancing compression efficiency without compromising key image details.
- **Comprehensive Evaluation and Benchmarking**: Extensive experiments are conducted to compare VCSC with traditional image compression techniques (JPEG, Huffman coding) and text-centered semantic communication methods. The results demonstrate that VCSC consistently achieves superior performance in structural similarity, perceptual quality, and robustness against wireless channel impairments.

The remainder of this paper is organized as follows. Section II reviews advancements in semantic communication and learned image compression. Section III describes the proposed VCSC. Section IV presents the proposed VCSC framework with quantization. Section V presents experimental evaluations, while Section V-D analyses the impact of design choices. Finally, Section VI concludes the paper and discusses future directions.

II. Related Work

In recent years, semantic communication has revolutionized wireless transmission by prioritizing the preservation of essential meaning over bit-precision accuracy [5], [6]. This paradigm shift involves encoding and decoding semantic content directly, rather than treating an image merely as a collection of symbols devoid of inherent semantics [7]. In [8], the authors introduced a semantic communication model that integrates a dynamic decision generation network and a generative adversarial network, effectively reducing required bandwidth while maintaining critical task-related information. In addition, Han *et al.* [9] explored the integration of generative models in semantic communication systems to enhance transmission efficiency and robustness. The aforementioned studies underscored that encoding the latent semantic content of an image offers greater resilience to noise and relaxes the bandwidth requirement in comparison to traditional pipeline-based codecs.

A. Semantic Communication and Image Transmission

Early studies on semantic communication emphasized transmitting meaning beyond symbol-level accuracy and were mainly discussed in the context of text or symbolic messages [10]. Building on this concept, an important line of research has developed vision-centered pipelines that follow an image–feature–image paradigm, where compact latent embeddings are transmitted to preserve both high-level semantics and local structures [11]. Representative works include end-to-end deep joint source–channel coding for semantic image transmission [12] and its attention-enhanced variants that prioritize informative latent regions to improve robustness under wireless impairments [13]. However, these methods are typically trained end-to-end for specific datasets and channel settings, and may require re-training when the data domain changes, which can limit generalization and increase the risk of overfitting, and this is a key concern for heterogeneous smart-city environments. More recently, the emergence of large pretrained models has enabled text-centered semantic communication, where images are converted into captions and the receiver re-synthesizes images from the transmitted text [14], [15]. While such pretrained priors can improve generalization, the image–text–image pipeline often loses fine-grained spatial details, making it less suitable for smart-city image transmission that demands faithful reconstruction. Therefore, we propose VCSC, which revisits the image–feature–image paradigm while leveraging a pretrained latent image representation to better balance generalization and spatial fidelity in smart-city deployments.

B. Learned Image Compression and Latent Space Models

The advent of deep learning has opened new avenues in learned image compression [16]. Pioneering works introduced

autoencoders that map images into compact latent code, surpassing classical codecs like JPEG in certain rate–distortion trade-offs [17]. Latent diffusion models have advanced the field by jointly learning semantic and structural features across multiple levels of granularity [18]. These latent models excel at capturing semantically relevant details even at significantly reduced spatial dimensions. For example, Jia *et al.* [19] explored the characteristics of latent space modeling in generative image compression, establishing a framework that effectively captures semantic content. Furthermore, Liu *et al.* [8] proposed a semantic communication model that integrates a dynamic decision generation network and a generative adversarial network, enhancing image reconstruction quality through adversarial and perceptual losses. These approaches are particularly valuable for real-time surveillance in crowded or bandwidth-limited scenarios.

C. Quantization Techniques for Semantic Image Transmission

Semantic image compression and transmission pipelines rely on quantization as a critical step to turn continuous deep features into compact discrete codes [20]. Early efforts often employed scalar quantizers that independently quantized individual coefficients [21]. With advancement of deep learning, researchers began integrating end-to-end differentiable quantization modules into convolutional or recurrent networks, ensuring that quantization errors are optimized for semantic relevance rather than mere pixel-level distortion [22]. Recent works extend this concept to scenarios with fluctuating channel conditions by dynamically adjusting quantization parameters based on channel feedback or error statistics [23]. Generative and adversarial models also employ sophisticated vector quantization strategies, enabling decoders to reconstruct high-quality images even when parts of the quantized code are lost or corrupted [24].

D. Semantic Robustness Under Noisy Channels

Smart city deployments frequently encounter dynamic interference and severe multipath fading, posing significant challenges for reliable image transmission [25], [26]. Traditional source and channel coding strategies often struggle to maintain image fidelity without increasing bit rates [27]. Recent studies have shown that coupling semantic encoding with adaptive channel modeling can mitigate these effects in resource-limited IoT devices [28], [29]. For example, Wang *et al.* [30] introduced a perceptual learned source-channel coding approach for high-fidelity image semantic transmission, combining encoder, wireless channel, decoder, and discriminator, which are jointly learned under both perceptual and adversarial losses, resulting in improved robustness against channel impairments. Similarly, Han *et al.* [9] proposed a generative model-based semantic communication approach that leverages GAN inversion methods to extract interpretable latent code, enhancing transmission efficiency and robustness. However, these methods are typically trained for specific channel settings, which limits their adaptability to the diverse and dynamic wireless conditions in smart-city environments.

E. Semantic Communication for Smart City Imaging

As smart cities continue to evolve, integrating semantic communication into urban imaging systems presents promising avenues for innovation [31]. One such advancement is federated learning, which enhances data trustworthiness and user participation in large-scale smart city sensing by enabling collaborative model training without centralized data collection, thereby addressing privacy and security concerns [32], [33]. Another advancement is the application of deep learning to efficient image transmission and analysis in traffic monitoring and infrastructure inspection, as it effectively captures complex data patterns [34]. Furthermore, the integration of edge computing with semantic communication frameworks allows for real-time data processing closer to the end users, which reduces latency and bandwidth usage, which is crucial for time-sensitive smart city applications [35], [36]. However, the capabilities of pretrained networks have not been fully explored in smart city applications.

III. VCSC System Description

As shown in Fig. 2, a VCSC system is a pretrained latent encoder-decoder network and is designed to exploit the latent space for efficient image transmission over wireless channels, aiming to preserve essential semantic information compared to the traditional TCSC counterpart. This VCSC system comprises two main parts: a transmitter and a receiver, connected by a wireless channel (i.e., AWGN or Rayleigh fading). Particularly, the transmitter is responsible for processing the input image from the camera and preparing it for transmission, which is composed of a semantic encoder and a quantization module. After processing at the transmitter, the image can be converted and represented by the quantized latent codes, which are fed into the wireless channel, where the quantized latent codes are transmitted. Without loss of generality and for simplicity, the wireless semantic transmission is assumed to occur under the AWGN or Rayleigh fading channel. After the wireless channel, the receiver aims to recover the transmitted image from these received quantized latent codes. Accordingly, the receiver comprises a dequantization module and a semantic decoder. The following subsections are detailed descriptions of each module of the system model under investigation.

A. Transmitter

1) Semantic Encoder: The semantic encoder transforms the input image into a compact latent code using the VAE encoder under the Stable Diffusion [18]. This pretrained encoder facilitates the extraction of semantically meaningful features while significantly reducing spatial redundancy. The encoding process comprises the following blocks:

i) *Convolutional Downsampling Block*: A convolutional layer followed by downsampling operations extracts low-level features and reduces the spatial resolution of the input image.

ii) *Residual Abstraction Block*: A stack of residual blocks refines the feature maps, capturing mid-level semantic information and improving representational depth.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549

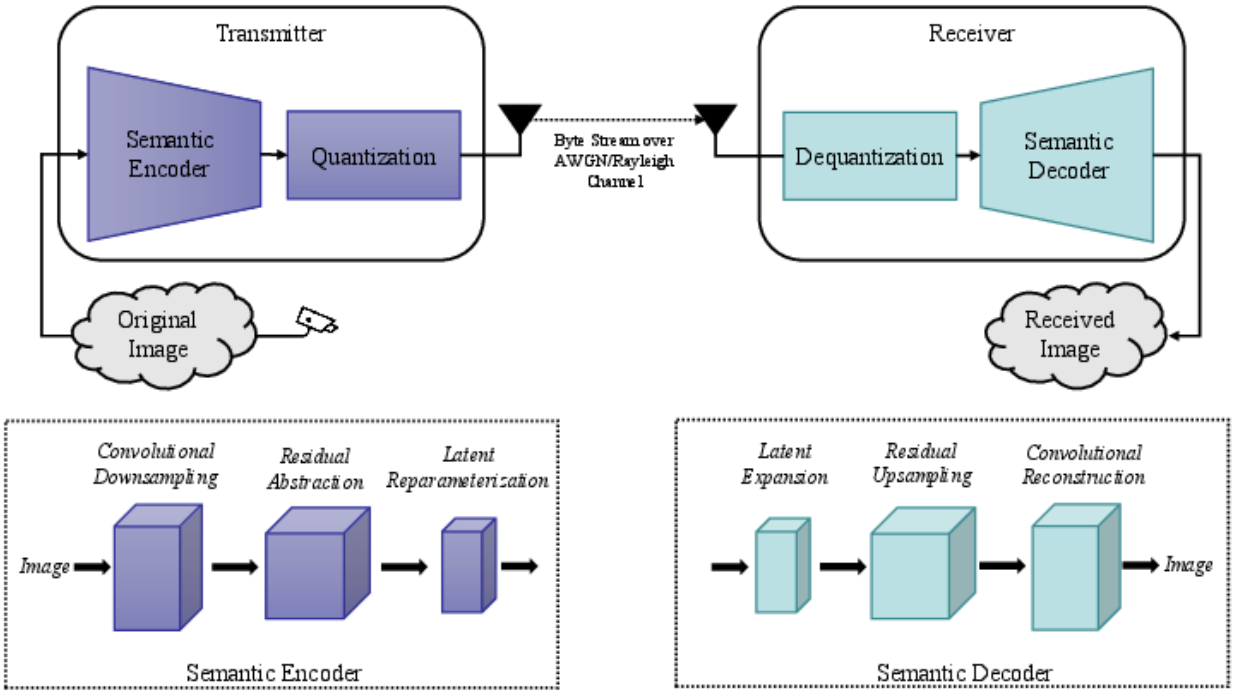


Fig. 2: The proposed VCSC system: the transmitter employs a semantic encoder and quantization modules to compress an image into a compact latent code, which is then transmitted over a noisy channel (e.g., AWGN or Rayleigh fading). At the receiver, dequantization and semantic decoding are applied to reconstruct the image.

iii) *Latent Reparameterization Block*: The encoder predicts the mean and variance of the latent distribution and applies the reparameterization trick to sample the latent code. This process yields a compact and expressive representation suitable for robust transmission.

2) Quantization Module: After generating the semantic representation, the encoder passes it to the quantization module, which plays a crucial role in preparing the semantic representation of the image for efficient transmission over the wireless channel.

The primary purpose of the quantization module is to reduce the amount of data that needs to be transmitted while ensuring that the most important information is preserved by converting the continuous or high-precision values of the semantic representation into discrete or lower-precision values. This process reduces the bit rate required for transmission, making the system more efficient. In this work, we investigate two quantization schemes, i.e., uniform 8-bit quantization and attention-based 4-bit quantization.

Specifically, the former is suitable for scenarios where a lower but acceptable precision is sufficient, and it divides the range of the semantic representation into uniform intervals, each represented by an 8-bit value. The number of intervals is determined by the quantization step size. The semantic representation values are then mapped to the nearest interval center, effectively reducing the precision of those regions. This ensures that even the less critical regions are quantized in a standardized and efficient manner, further contributing to the overall efficiency of the system. The latter leverages a spatial attention mechanism to identify and preserve the details of critical regions by allocating higher precision to them. Simultaneously, 4-bit quantization is applied to less critical regions, converting their values into discrete 4-bit representations. This approach ensures efficient utilization of bandwidth, while maintaining the essential semantic details of the image, and, in the meantime, enhances the robustness of the transmitted data against wireless channel impairments. This attention mechanism identifies critical regions within the semantic representation that contain significant semantic information, which is significant for preserving the overall meaning and quality of the image.

B. Wireless Channel Modeling

The quantized latent codes are transmitted over the wireless channel modeled as AWGN or Rayleigh fading. The VCSC system addresses these channel impairments by incorporating techniques to enhance the robustness of the transmitted signal. This includes the design of the quantization process and the integration of error correction mechanisms within the overall communication pipeline.

C. Receiver

1) Dequantization Module: The main goal of the dequantization module at the receiver is to reverse the quantization process that was applied at the transmitter. When the quantized latent codes are received, they undergo dequantization to transform them back into a form that is compatible with the semantic decoder. This critical step reconstructs the semantic representation of the image, which was compressed during quantization, preparing it for the final image reconstruction process. The dequantization process reconstructs the original

continuous values from discrete quantized data, or approximates them when exact recovery is infeasible. This minimizes semantic information lost during quantization, allowing for a more accurate recovery of the image details.

2) Semantic Decoder: The semantic decoder reconstructs the original image from the dequantized latent code using the Variational Autoencoder (VAE) decoder under the Stable Diffusion [18]. This pretrained decoder mirrors the encoder's operations, ensuring perceptually consistent image reconstruction. The decoding process involves the following blocks:

1) *Latent Expansion Block*: The dequantized latent code is projected into a higher-dimensional feature space through a learnable transformation.
2) *Residual Upsampling Block*: A combination of upsampling layers and residual blocks progressively restores spatial resolution and enhances semantic consistency.
3) *Convolutional Reconstruction Block*: A final convolutional layer transforms the refined feature maps into the RGB image space, yielding the reconstructed output image.

IV. Pretrained Network for Vision-Centered Semantic Transmission Framework with Quantization

This section investigates the pretrained latent encoder-decoder network with a quantization-dequantization mechanism for the VCSC system. Before delving into the details, we first summarize the notations in Table I, providing a concise overview of the mathematical symbols introduced in this work.

TABLE I: Summary of Notations

Symbol	Description
$\mathbf{x}$	Input image.
$\hat{\mathbf{x}}$	Reconstructed image.
D_{sem}	Semantic distortion.
$\mathcal{S}$	Similarity function between $\mathbf{x}$ and $\hat{\mathbf{x}}$.
$\mathbf{z}$	Latent code of image.
$\boldsymbol{\mu}(\mathbf{x})$	Mean vector.
$\boldsymbol{\sigma}(\mathbf{x})$	Standard deviation vector.
$\boldsymbol{\epsilon}$	Gaussian noise vector $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
q	Quantized latent code.
$\tilde{q}$	Received quantized code after channel noise.
$\tilde{\mathbf{z}}$	Dequantized latent code at the receiver.
$q_{\text{high}}, q_{\text{low}}$	Quantized values by region importance (high vs. low).
q_{com}	Combined quantized latent code with attention.
$\tilde{q}_{\text{com}}$	Received combined quantized code with attention.
$s, s_{\text{high}}, s_{\text{low}}$	Scale factors for quantization.
$z_{\min}, z_{\max}$	Min and max values for uniform quantization.
$z_{\min}^{\text{high}}, z_{\max}^{\text{high}}$	Min and max values in high-attention regions.
$z_{\min}^{\text{low}}, z_{\max}^{\text{low}}$	Min and max values in low-attention regions.
C	Number of channels in the latent code.
$A(i,j)$	Attention value at spatial location (i,j).
$A_{\text{norm}}(i,j)$	Normalized attention value.
$M(i,j)$	Binary attention mask at location (i,j).
T	Mask threshold.
$Loss$	Loss function.
λ	Weighting (trade-off) coefficient.
ν^2	Noise variance in AWGN channel.
$h(t)$	Time-varying Rayleigh fading coefficient.
f_D	Doppler frequency.
θ_n, ϕ_n	Angle and phase used in Rayleigh channel model.
η	Additive Gaussian noise in AWGN, $\eta \sim \mathcal{N}(0, \nu^2)$.
SNR	Signal-to-noise ratio.

A. Definition of Visual Semantics

In this work, we define visual semantics as the preservation of structural and semantic integrity in the latent space. Formally, given an input image $\mathbf{x}$ and reconstructed image $\hat{\mathbf{x}}$, semantic distortion D_{sem} can be expressed as:

$$D_{\text{sem}} = 1 - \mathcal{S}(\mathbf{x}, \hat{\mathbf{x}}), \tag{1}$$

where $\mathcal{S}$ denotes the similarity between the input image $\mathbf{x}$ and reconstructed image $\hat{\mathbf{x}}$. Hence, our objective is to minimize both structural and semantic distortion simultaneously. In practice, $\mathcal{S}$ can be instantiated using perceptual or semantic similarity metrics, including the Structural Similarity Index (SSIM) [37], Contrastive Language-Image Pretraining (CLIP) score [38], and Learned Perceptual Image Patch Similarity (LPIPS) [39]. Additionally, the Frechet Inception Distance (FID) [40] is employed to evaluate distributional alignment between the reconstructed and original images, complementing the semantic distortion measure.

B. Semantic Encoder of Transmitter with Quantization

1) Semantic Encoder: To extract the semantics of an image, the semantic encoder processes an input image $\mathbf{x}$ and produces a latent code $\mathbf{z} \in \mathbb{R}^{4\times 64\times 64}$. To enable stochastic sampling of the latent code while preserving gradient flow during training, the semantic encoder predicts the mean $\boldsymbol{\mu}(\mathbf{x})$ and log-variance $\log \boldsymbol{\sigma}^2(\mathbf{x})$ of the latent distribution, and samples $\mathbf{z}$ as:

$$\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\boldsymbol{\sigma}(\mathbf{x}) = \exp\left(\frac{1}{2}\log \boldsymbol{\sigma}^2(\mathbf{x})\right)$, and $\odot$ denotes element-wise multiplication. The random noise vector $\boldsymbol{\epsilon}$ is drawn from a standard Gaussian distribution. This formulation allows the model to learn a smooth and expressive latent space for semantic communication.

2) Quantization: To enable efficient transmission of the latent code $\mathbf{z}$, two quantization schemes are proposed to balance compression efficiency and semantic fidelity, as described in a) and b) below. Specifically, the uniform 8-bit quantization is used when bandwidth is less constrained to prioritize reconstruction fidelity, while the attention-based 4-bit quantization is designed for strict bandwidth-limited scenarios by allocating more bits to semantically important regions.

a) Uniform 8-bit Quantization: The latent code is quantized uniformly to an 8-bit representation. First, the global minimum and maximum values of $\mathbf{z}$ are computed as:

$$z_{\min} = \min(\mathbf{z}), \quad z_{\max} = \max(\mathbf{z}), \tag{3}$$

Based on these extrema, the scale factor s is determined as:

$$s = \frac{255}{z_{\max} - z_{\min}}, \tag{4}$$

Using $z_{\min}$ and $z_{\max}$ defined in Eq. 3, the latent code is then quantized by applying rounding and clipping within the valid range, as follows:

$$q = \text{clip}\Big(\text{round}\big((\mathbf{z} - z_{\min}) \cdot s\big),\, 0,\, 255\Big), \tag{5}$$

The uniform 8-bit quantization ensures efficient encoding while maintaining the essential semantic structure of the latent code.

b) Attention-Based 4-bit Quantization: To further enhance compression efficiency while preserving semantic content, an alternative adaptive quantization scheme based on an attention mechanism is proposed. This approach dynamically adjusts the quantization precision based on the importance of different spatial regions in the latent code.

First, an attention map A is computed to measure the relative importance of each spatial location (i, j) across all channels of $\mathbf{z}$, as follows:

$$A(i,j) = \frac{1}{C}\sum_{c=1}^{C} |\mathbf{z}_{c,i,j}|\,, \tag{6}$$

where C is the number of channels, and $\mathbf{z}_{c,i,j}$ represents the latent code at channel c and spatial location (i, j). This operation calculates the mean absolute activation over all channels, highlighting regions with stronger responses. The choice of mean absolute activation in Eq. 6 provides a computationally efficient and stable measure of spatial importance. Regions with stronger activations typically correspond to semantically salient objects or structures in the scene.

Next, the attention map A in Eq. 6 is normalized to a range of $[0, 1]$ to facilitate thresholding, as follows:

$$A_{\text{norm}}(i,j) = \frac{A(i,j) - \min(A)}{\max(A) - \min(A)}, \tag{7}$$

Based on the normalized attention map A_{norm}, computed from A in Eq. 6, a binary mask M is generated using a predefined threshold T, as follows:

$$M(i,j) = \begin{cases} 1, & \text{if } A_{\text{norm}}(i,j) \geq T, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where $M(i, j)$ identifies important regions ($M(i, j) = 1$), which are assigned a higher quantization precision, while less significant regions ($M(i, j) = 0$) are assigned lower precision. The threshold T is a tunable hyperparameter that determines which spatial regions are assigned higher quantization precision. In this study, T was empirically set to 0.2 and 0.4, as these values yielded a good balance between compression rate and semantic fidelity during our experiments.

Two separate quantization schemes are then applied as follows (see Eqs. 9 and 11):

High-precision (8-bit) quantization for important regions:

$$q_{\text{high}} = \text{clip}\Big(\text{round}\Big((\mathbf{z} - z_{\min}^{\text{high}}) \cdot s_{\text{high}}\Big),\, 0,\, 255\Big), \tag{9}$$

where $z_{\min}^{\text{high}}$ and $z_{\max}^{\text{high}}$ represent the minimum and maximum values of $\mathbf{z}$ in high-importance regions, and the scale factor s_{high} is defined as:

$$s_{\text{high}} = \frac{255}{z_{\max}^{\text{high}} - z_{\min}^{\text{high}}}, \tag{10}$$

Low-precision (4-bit) quantization for less important regions:

$$q_{\text{low}} = \text{clip}\Big(\text{round}\Big((\mathbf{z} - z_{\min}^{\text{low}}) \cdot s_{\text{low}}\Big),\, 0,\, 15\Big), \tag{11}$$

where $z_{\min}^{\text{low}}$ and $z_{\max}^{\text{low}}$ are the corresponding extrema for low-importance regions, and the scale factor s_{low} is:

$$s_{\text{low}} = \frac{15}{z_{\max}^{\text{low}} - z_{\min}^{\text{low}}}, \tag{12}$$

Finally, the two quantized representations are merged based on the mask M:

$$q_{\text{com}}(i,j) = M(i,j)\, q_{\text{high}}(i,j) + \big(1 - M(i,j)\big)\, q_{\text{low}}(i,j), \tag{13}$$

The attention-based 4-bit quantization ensures that important regions retain high fidelity, while less critical regions are compressed more aggressively, thereby achieving an effective trade-off between compression rate and semantic preservation. The current bit allocation strategy relies on min–max scaling of activations within high- and low-attention regions, which provides computational simplicity and efficiency. This design is motivated by real-time transmission constraints in smart city scenarios.

C. Receiver

At the receiver, the transmitted bitstream is first recovered using channel decoding and demodulation, resulting in the received quantized latent code $\tilde{q}$ (or $\tilde{q}_{\text{com}}$ for attention-based 4-bit quantization). The variable $\tilde{q}$ is the noisy version of the transmitted code q after passing through the communication channel, as described in detail in Section IV-E. The receiver then reconstructs the latent code and ultimately the image as follows.

1) Dequantization:

a) Uniform 8-bit Dequantization: For uniform quantization, the latent code is recovered by inverting the quantization process:

$$\tilde{\mathbf{z}} = \frac{\tilde{q}}{s} + z_{\min}, \tag{14}$$

where $\tilde{q}$ is the received quantized latent code after transmission through the channel, $s = \frac{255}{z_{\max} - z_{\min}}$ is the quantization scale factor, and $z_{\min}$ is the global minimum of the latent values.

b) Attention-Based 4-bit Dequantization: When an attention-based 4-bit quantization is used, the recovered latent code is computed according to the binary mask M:

$$\tilde{\mathbf{z}}(i,j) = \begin{cases} \frac{\tilde{q}_{\text{com}}(i,j)}{s_{\text{high}}} + z_{\min}^{\text{high}}, & \text{if } M(i,j) = 1, \\ \frac{\tilde{q}_{\text{com}}(i,j)}{s_{\text{low}}} + z_{\min}^{\text{low}}, & \text{if } M(i,j) = 0, \end{cases} \tag{15}$$

where $\tilde{q}_{\text{com}}(i, j)$ denotes the received attention-guided quantized value at position (i, j), s_{high} and s_{low} are the scale factors for high- and low-importance regions, respectively, and $z_{\min}^{\text{high}}$ and $z_{\min}^{\text{low}}$ are their corresponding minimum values used during quantization.

2) Semantic Decoder: The semantic decoder reconstructs the output image $\hat{\mathbf{x}}$ from the dequantized latent code $\tilde{\mathbf{z}}$, mirroring the architecture of the semantic encoder. The semantic decoder generates the reconstructed image:

$$\hat{\mathbf{x}} = D(\tilde{\mathbf{z}}), \tag{16}$$

where $D(\cdot)$ denotes the pretrained decoder function.

D. Pretrained Encoder–Decoder

The semantic encoder and decoder of VCSC are adopted from the Stable Diffusion [18], which has been jointly pretrained on the large-scale LAION dataset, comprising approximately 400 million image-text pairs. This pretrained network provides a robust mapping between images and a lower-dimensional latent space, endowing the system with strong initial semantic representation capabilities. The loss function used during pretraining is defined as:

$$Loss = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda \, \mathrm{LPIPS}(\mathbf{x}, \hat{\mathbf{x}}), \tag{17}$$

where $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ is the reconstruction loss computed as the squared Euclidean distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$, $\mathrm{LPIPS}(\mathbf{x}, \hat{\mathbf{x}})$ is the perceptual loss that measures semantic similarity via deep feature comparisons [39], and λ is a hyperparameter balancing the contribution of the perceptual loss relative to the reconstruction loss.

E. Channel Model

The quantized data, represented as q (or q_{com} in the attention-based scheme), is serialized into a byte stream and transmitted over a noisy channel. To emulate realistic transmission conditions, two channel models are considered: AWGN and Rayleigh fading.

The AWGN channel is modeled as:

$$\tilde{q} = q + \eta, \quad \eta \sim \mathcal{N}(0, \nu^2), \tag{18}$$

where q is the transmitted quantized latent code, $\tilde{q}$ is the received noisy version, and η is additive Gaussian noise with variance ν^2, determined by the target Signal-to-Noise Ratio (SNR). The received code $\tilde{q}$ is then used for dequantization at the receiver.

In environments such as smart cities, the channel is subject to multipath fading, which is represented as a time-varying Rayleigh fading channel with additive noise:

$$\tilde{q}(t) = h(t) \cdot q(t) + \eta(t), \quad \eta(t) \sim \mathcal{N}(0, \nu^2), \tag{19}$$

where $h(t)$ represents the time-varying fading coefficient. In the implementation, $h(t)$ is generated using a Jakes model:

$$h(t) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \exp\Big\{ j \Big(2\pi f_D \, t \cos\theta_n + \phi_n \Big) \Big\}. \tag{20}$$

with:

- N being the number of scatterers (e.g., 16),
- f_D the Doppler frequency (e.g., 10 Hz),
- θ_n uniformly distributed over $[0, 2\pi)$, and
- ϕ_n random phases uniformly distributed in $[0, 2\pi)$.

This formulation yields a Rayleigh fading coefficient with the appropriate time-variation, and the received quantized code $\tilde{q}(t)$ is used for dequantization and subsequent semantic decoding. It is worth noting that channel noise is injected only during the simulation stage to emulate realistic transmission conditions, rather than in the training phase of the encoder–decoder. This is consistent with the design of VAE, where the encoder already introduces stochasticity via the reparameterization trick.

V. Numerical Results

In this section, the performance of the proposed VCSC is compared with several baseline methods, including Huffman coding, JPEG compression, and text-centered semantic communication, under both AWGN and Rayleigh fading channels. The experiments were conducted on a system equipped with an Intel Core i7-9700 CPU and an NVIDIA GeForce RTX 4090 GPU.

A. Simulation Settings

1) Proposed VCSC: The encoder and decoder of the proposed VCSC adopt the pretrained model of "stable-diffusion-v1-5" [18]. The input image is first converted into a latent code (of size $1 \times 4 \times 64 \times 64$) using the encoder. For quantization, VCSC employs the uniform 8-bit quantization (8-bit), and an alternative version with the attention-based 4-bit quantization (4-bit-attn) is also evaluated. The quantized latent code is then converted into a byte stream, transmitted over a communication channel, and subsequently dequantized at the receiver. Finally, the latent code is decoded back into an image using the decoder.

2) Baseline Methods: For comparison, three alternative methods were considered:

- Huffman. The original image is converted into a byte stream and compressed using Huffman coding. The compressed data is then transmitted over the communication channel and decoded to recover the image.
- JPEG. JPEG compression is applied to the image using a predetermined quality factor Q, which is set to 80 in the experiments because this value typically offers a good balance between compression efficiency and image quality. The resulting JPEG-compressed byte stream is transmitted and decoded to reconstruct the image.
- Image-Text-Image Semantic Communication (ItISC). ItISC was implemented to simulate the text-centered semantic communication paradigm for image transmission. In ItISC, the original image is first processed by a pretrained Bootstrapping Language-Image Pretraining model (BLIP) [41] to generate a textual descriptive caption capturing the key semantic elements of the scene. This caption, which encapsulates objects, actions, and contextual relationships, is then encoded into a UTF-8 byte stream and transmitted. At the receiver, the transmitted bit stream is demodulated and decoded to recover the text. The recovered text is subsequently fed into a Stable Diffusion model [18] text-to-image pipeline, which generates an image that semantically corresponds to the original input. Note that ItISC re-synthesizes an image conditioned only on the transmitted caption, so it is not expected to preserve low-level textures or exact spatial details of the original image; therefore, weaker LPIPS/FID scores are anticipated even when semantic consistency is maintained.

For fairness, all baseline methods (JPEG, Huffman, and ItISC) are transmitted with Turbo coding enabled, consistent with the proposed VCSC. We do not include the semantic image transmission and generative schemes in the literature [11]–[13]

as quantitative baselines, since they are trained for a specific dataset/channel setting, whereas our VCSC operates in the frozen latent space of a pretrained diffusion model and targets generalization across heterogeneous smart-city data without fine-tuning.

3) Experimental Settings: All experiments are conducted under a unified simulation framework. Both AWGN and Rayleigh fading channels are simulated to assess performance under varying channel conditions. Transmission is performed using BPSK modulation/demodulation at a specified SNR (in dB, where a higher SNR indicates lower noise power), with additional tests conducted over a range of SNR values. To enhance error resilience, Turbo coding is applied for channel encoding and decoding in evaluated methods. Its implementation follows the approach in [42], utilizing the log-MAP algorithm with 5 iterations.

B. Dataset Description and Evaluation Metrics

In this study, the proposed method was evaluated using the Traffic Detection Project Dataset [43] and the SmartCity Dataset [44]. The Traffic Detection Project Dataset offers a rich collection of traffic camera images from various countries, providing diverse geographic coverage for global traffic monitoring. Captured under diverse weather, lighting, and traffic conditions, the dataset reflects real-world challenges. It comprises 5805 training images and 279 test images. The SmartCity Dataset was collected independently, comprising a total of 50 images from 10 different urban environments, such as office entrances, sidewalks, atriums, and shopping malls. These images are all captured from a high-angle perspective, mimicking typical video surveillance scenarios. This dataset is intentionally designed to include both indoor and outdoor scenes with relatively few pedestrians.

To quantitatively assess VCSC, four evaluation metrics are employed: SSIM [37], CLIP [38], LPIPS [39], and FID [40].

SSIM score evaluates the structural similarity between the original image $\mathbf{x}$ and the reconstructed image $\hat{\mathbf{x}}$ by considering luminance, contrast, and structure. It is defined as:

$$\text{SSIM Score} = \frac{(2\bar{p}_{\mathbf{x}}\bar{p}_{\hat{\mathbf{x}}} + \kappa_1)(2\varsigma_{p_{\mathbf{x}\hat{\mathbf{x}}}} + \kappa_2)}{(\bar{p}_{\mathbf{x}}^2 + \bar{p}_{\hat{\mathbf{x}}}^2 + \kappa_1)(\varsigma_{p_{\mathbf{x}}}^2 + \varsigma_{p_{\hat{\mathbf{x}}}}^2 + \kappa_2)}, \tag{21}$$

where $\bar{p}_{\mathbf{x}}$ and $\bar{p}_{\hat{\mathbf{x}}}$ are the mean pixel values (i.e., average intensity over all pixels in the image), $\varsigma_{p_{\mathbf{x}}}$ and $\varsigma_{p_{\hat{\mathbf{x}}}}$ are the corresponding standard deviations, and $\varsigma_{p_{\mathbf{x}\hat{\mathbf{x}}}}$ is the cross-covariance between the two pixel-level images. The constants $\kappa_1 = (K_1 L)^2$ and $\kappa_2 = (K_2 L)^2$ stabilize the division, where L is the dynamic range, and K_1, K_2 are small constants.

CLIP score measures the semantic similarity between the original image and the reconstructed image in the CLIP embedding space by computing:

$$\text{CLIP Score} = \cos(\mathbf{x}_e, \hat{\mathbf{x}}_e), \tag{22}$$

where $\mathbf{x}_e$ and $\hat{\mathbf{x}}_e$ are the visual embeddings of the original and reconstructed images, respectively.

LPIPS score evaluates the perceptual similarity between the original and reconstructed images by comparing deep features extracted from neural networks:

$$\text{LPIPS Score} = \sum_l \frac{1}{W_l H_l} \sum_{h,w} \|w_l \odot (f_{hw}^l(\mathbf{x}) - f_{hw}^l(\hat{\mathbf{x}}))\|_2^2, \tag{23}$$

where $f_{hw}^l(\cdot)$ denotes the activation at location (h, w) in layer l, and w_l is a weighting vector.

FID score measures the similarity between the distributions of original and reconstructed images by comparing their feature representations extracted from a pretrained Inception network. It is defined as:

$$\text{FID Score} = \|\bar{\chi} - \bar{\psi}\|_2^2 + \text{Tr}\left[\Sigma_\chi + \Sigma_\psi - 2\left(\Sigma_\chi \Sigma_\psi\right)^{\frac{1}{2}}\right], \tag{24}$$

where $\bar{\chi}$ and $\bar{\psi}$ are the mean feature vectors, and Σ_χ, Σ_ψ are the covariance matrices computed from the original image $\mathbf{x}$ and reconstructed image $\hat{\mathbf{x}}$, respectively.

For SSIM and CLIP, higher values indicate better structural/semantic consistency, whereas for LPIPS and FID, lower values are better as they measure perceptual distance and distributional discrepancy, respectively.

C. Evaluation Results of VCSC and Various Methods

1) On the Traffic Detection Project Dataset: Fig. 3 compares the proposed VCSC (8-bit) against three baseline approaches (Huffman, JPEG, and ItISC) under varying SNR conditions in a communication channel for the Traffic Detection Project Dataset.

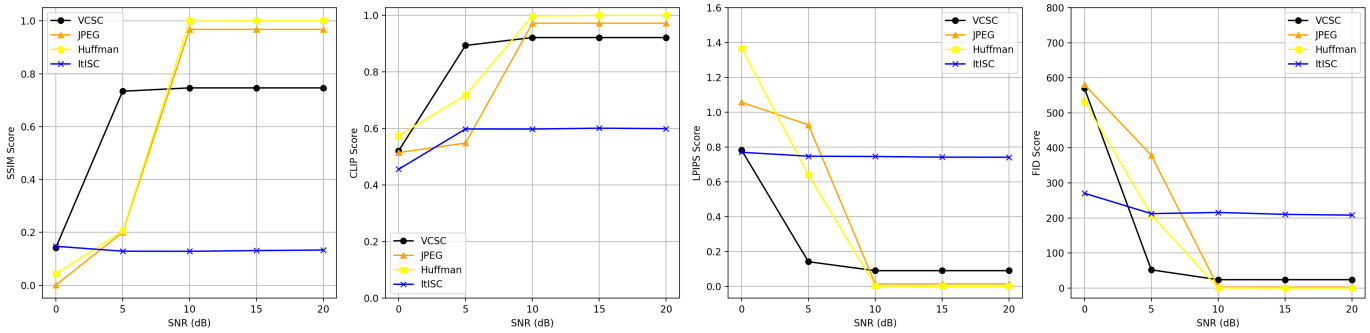
Fig. 3a shows the results under the AWGN channel. Performance trends emerge as the SNR increases from 0 dB to 25 dB:

- **SSIM Score.** VCSC rapidly increases from above 0.18 at 0 dB to above 0.75 by 5 dB, stabilizing around 0.78 beyond 10 dB. While Huffman and JPEG eventually reach about 0.98 at 10 dB, their initial performance is weaker. In contrast, ItISC consistently remains significantly lower (around 0.15) across all SNR levels.
- **CLIP Score.** VCSC starts at above 0.5 at 0 dB, quickly surpasses above 0.9 by 10 dB and remains near perfect at higher SNRs. Huffman and JPEG reach a peak of nearly 1.0, but only after 10 dB. ItISC stabilizes around 0.6.
- **LPIPS Score.** VCSC starts at about 0.78 at 0 dB (lower than Huffman at about 1.35 and JPEG at about 1.1), and drops sharply to near 0.1 by 10 dB, maintaining stability thereafter. ItISC stagnates above 0.75, failing to improve perceptual fidelity.
- **FID Score.** VCSC begins at about 550, and rapidly decreases to near 40 by 10 dB, outperforming all other methods (Huffman, JPEG, and ItISC). ItISC remains above 200, failing to enhance image quality at higher SNRs.

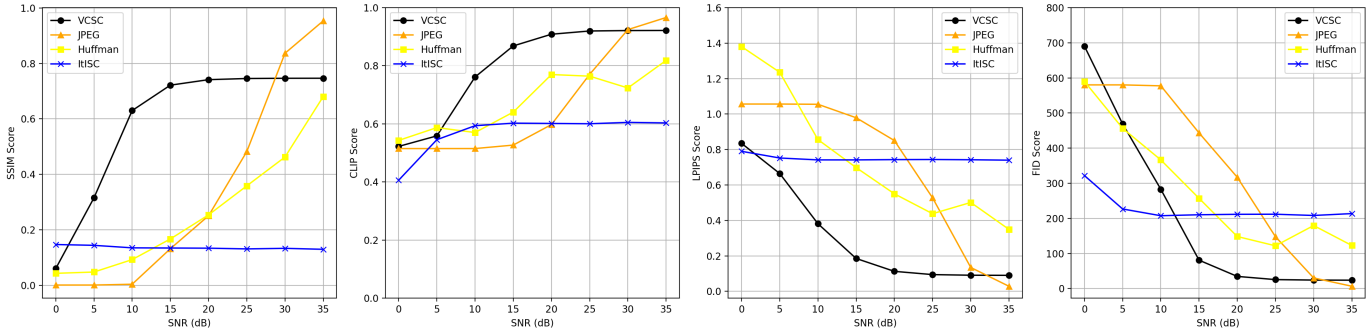
Fig. 3b shows the results under the Rayleigh fading channel. Performance trends emerge as the SNR increases from 0 dB to 35 dB:

- **SSIM Score.** VCSC rapidly increases from about 0.1 at 0 dB to above 0.75 by 15 dB, stabilizing around 0.78 at

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549

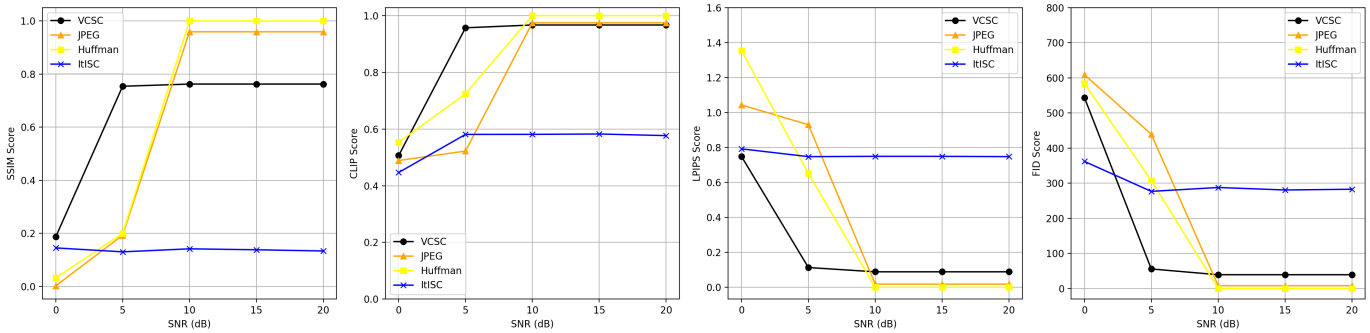


(a) AWGN channel.

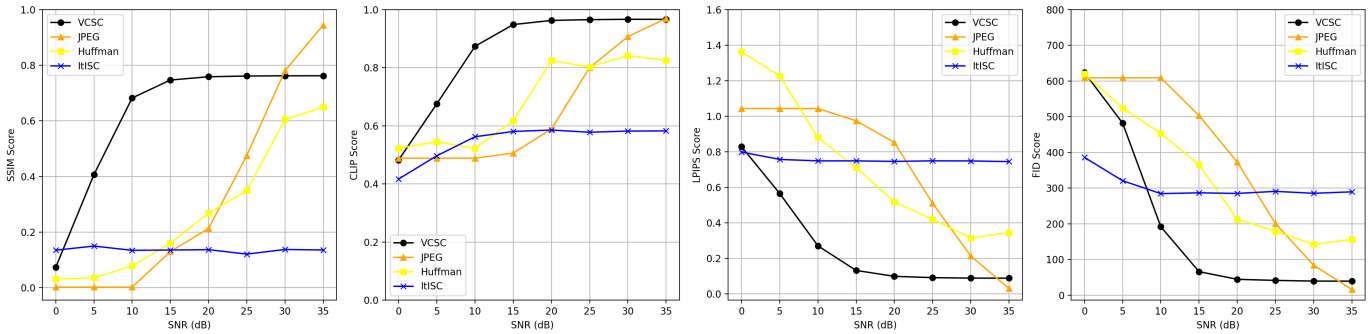


(b) Rayleigh fading channel.

Fig. 3: Performance comparison of VCSC and baseline methods (Huffman, JPEG, ItISC) on the Traffic Detection Project Dataset under (a) AWGN and (b) Rayleigh fading channels across varying SNR levels.



(a) AWGN channel.



(b) Rayleigh fading channel.

Fig. 4: Performance comparison of VCSC and baseline methods (Huffman, JPEG, ItISC) on the SmartCity Dataset under (a) AWGN and (b) Rayleigh fading channels across varying SNR levels.

25 dB. In contrast, Huffman and JPEG show slower improvements, and Huffman only surpasses VCSC beyond 30 dB, while ItISC remains consistently below 0.2.

- **CLIP Score.** VCSC starts at about 0.45 at 0 dB, surpasses 0.9 by 20 dB, and maintains near-perfect scores at higher SNR levels. Huffman and JPEG improve gradually, reaching about 0.98 and about 0.78 respectively, while ItISC remains stagnant around 0.6.
- **LPIPS Score.** VCSC starts at above 0.8 at 0 dB, lower than Huffman (about 1.4) and JPEG (about 1.05), and drops sharply to near 0.1 by 20 dB. ItISC remains above 0.75.
- **FID Score.** VCSC begins at about 700 at 0 dB and rapidly decreases to near 40 by 20 dB, outperforming competing methods (Huffman, JPEG, and ItISC). Meanwhile, ItISC remains above 200, struggling to enhance image quality.

2) On the SmartCity Dataset: Fig. 4 presents a comparison of the proposed VCSC (8-bit) against three baseline methods (Huffman, JPEG, and ItISC) over a range of SNR values for the SmartCity Dataset.

Fig. 4a presents the results for the AWGN channel. Performance trends emerge as the SNR increases from 0 dB to 25 dB:

- **SSIM Score.** VCSC rapidly rises above 0.75 at 5 dB and stabilizes near 0.78 by 10 dB, maintaining its advantage over ItISC throughout. Although Huffman and JPEG reach 0.98 at 20 dB, VCSC achieves high perceptual quality much earlier, highlighting its robustness under challenging channel conditions.
- **CLIP Score.** While Huffman and JPEG peak at 1.0 by 10 dB, VCSC closely follows at 0.98, demonstrating superior consistency across noisy conditions. ItISC, in contrast, lags behind at 0.6. The results confirm that VCSC achieves high semantic fidelity while balancing robustness in varying SNR conditions.
- **LPIPS Score.** VCSC starts at about 0.75 (lower than Huffman/JPEG$>$1.0) at 0 dB, and quickly drops to about 0.1 by 10 dB, maintaining high visual quality across all SNR levels. In contrast, ItISC remains high (about 0.8), failing to improve significantly.
- **FID Score.** VCSC begins at around 560 (lower than Huffman/JPEG at about 600), and drops sharply to around 30 by 10 dB, matching Huffman while significantly outperforming ItISC (nearly 300 at high SNR).

Fig. 4b presents the results for the Rayleigh fading channel. Performance trends emerge as the SNR increases from 0 dB to 35 dB:

- **SSIM Score.** VCSC starts at about 0.1 at 0 dB, rapidly increases beyond 0.75 by 10 dB, and stabilizes around 0.78 by 20 dB. In contrast, Huffman and JPEG improve gradually, while ItISC remains stagnant below 0.2, highlighting VCSC's superior adaptability.
- **CLIP Score.** VCSC starts at about 0.45 at 0 dB and surpasses 0.95 by 20 dB. Huffman and JPEG peak at 0.85 and 0.95 respectively, while ItISC remains lower at 0.6, confirming VCSC's robustness in maintaining semantic content.
- **LPIPS Score.** VCSC starts lower than Huffman (about 1.35) and JPEG (about 1.05) at 0 dB and drops sharply to near 0.1 by 20 dB. Huffman and JPEG improve slowly, while ItISC stagnates at approximately 0.8, demonstrating its inefficiency.
- **FID Score.** VCSC begins at around 620, and plummets to nearly 50 by 20 dB, outperforming all benchmarks (Huffman, JPEG, and ItISC). In contrast, ItISC remains above 300, indicating persistent degradation in high-SNR conditions.

3) Comparison Results of VCSC between 8-bit and 4-bit-attn: Fig. 5 compares the performance of the proposed VCSC under different quantization settings: VCSC (8-bit) for 8-bit uniform quantization, and VCSC (4-bit-attn-0.2) and VCSC (4-bit-attn-0.4) for 4-bit attention quantization with thresholds of 0.2 and 0.4, respectively. The results are compared against JPEG on the SmartCity Dataset. The results are presented for both the AWGN channel (Fig. 5a) and the Rayleigh fading channel (Fig. 5b), illustrating how each method performs under varying levels of channel noise.

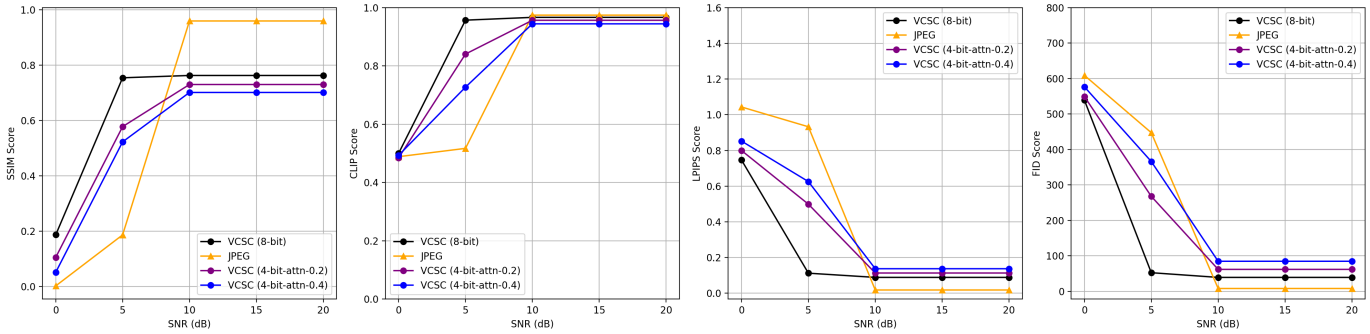
Under the AWGN channel, VCSC (8-bit) achieves the highest SSIM and CLIP Score, demonstrating superior structural and semantic fidelity. VCSC (4-bit-attn-0.2) and VCSC (4-bit-attn-0.4) closely follow, especially at moderate to high SNRs, while benefiting from reduced bit usage. VCSC (4-bit-attn-0.2) tends to slightly outperform VCSC (4-bit-attn-0.4) in preserving image quality, as reflected by higher SSIM and CLIP scores and lower LPIPS and FID scores. In contrast, JPEG shows a pronounced drop in performance at lower SNRs, indicating lower resilience to noise.

Under the Rayleigh fading channel, VCSC (8-bit) achieves the highest overall reconstruction quality, while VCSC (4-bit-attn-0.2) and VCSC (4-bit-attn-0.4) exhibit comparable performance in the mid to high SNR ranges, reinforcing the effectiveness of attention-based quantization. Despite JPEG's competitive performance at higher SNR values, its image reconstruction quality declines when facing harsh fading conditions.

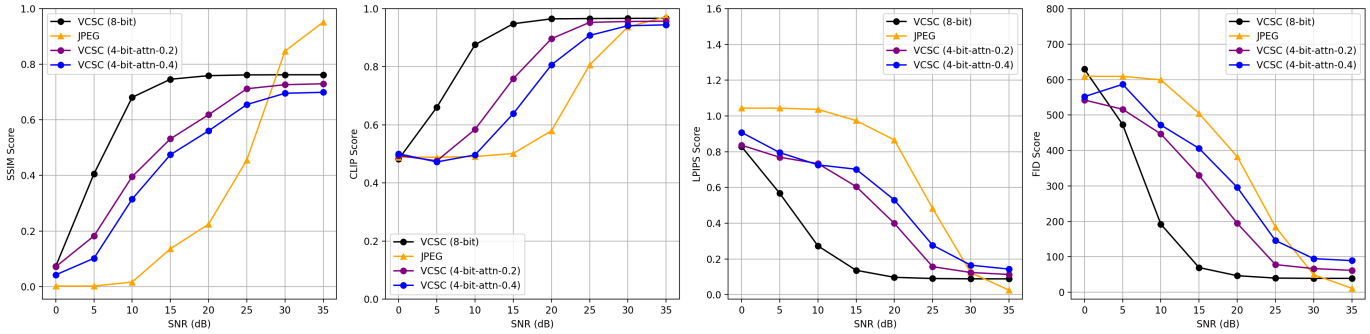
4) The Comparison Results of the Compression Rates: Table II presents the compression rates for different methods evaluated on the SmartCity Dataset, where the compression rate is computed as the ratio between the compressed data size (in bits) and the original image size (in bits). The table clearly illustrates that VCSC methods achieve high compression efficiency, while Huffman and JPEG exhibit low compression efficiency. Although ItISC applies aggressive compression, its capacity to retain sufficient image details is limited. These results highlight that the proposed VCSC offers a balanced trade-off between compression efficiency and image quality. Overall, the findings underscore the effectiveness of VCSC in achieving robust image transmission with minimal bit usage, especially under challenging wireless channel conditions.

5) Computational Overhead and Deployment Scenario: Our intended deployment scenario follows a typical smart-city pipeline [45], where resource-constrained cameras act as transmitters and a data center performs the receiver-side reconstruction. Table III reports the average per-image runtime of different schemes. Compared with traditional compression

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549



(a) AWGN channel.



(b) Rayleigh fading channel.

Fig. 5: Comparison of VCSC with 8-bit uniform and 4-bit attention-based quantization (thresholds 0.2 and 0.4), evaluated on the SmartCity Dataset under (a) AWGN and (b) Rayleigh fading channels. JPEG is included as a baseline.

TABLE II: Compression rates of VCSC and baselines, where the compression rate is computed as $\frac{\text{Compressed Bits}}{\text{Original Bits}}$. A smaller value means stronger compression.

Method	Compression Rate
ItISC	0.01%
JPEG	6.5%
Huffman	95.4%
VCSC (8-bit)	2.1%
VCSC (4-bit-attn-0.4)	1.7%
VCSC (4-bit-attn-0.2)	1.5%

TABLE III: Average per-image runtime of different methods on the SmartCity Dataset.

Method	Transmitter (ms)	Receiver (ms)
JPEG	1.8	2.3
ItISC	392.5	4342.1
Huffman	1031.2	2010.7
VCSC (8-bit)	93.8	188.7
VCSC (4-bit-attn)	94.9	187.5

(e.g., JPEG), VCSC introduces additional computation at the transmitter due to semantic encoding, requiring about *93–95* ms per image for both the 8-bit and the attention-guided 4-bit variants. Meanwhile, the receiver-side overhead is *187–189* ms per image, which is lightweight for data center deployment. Notably, despite this added encoding cost relative to JPEG, VCSC remains substantially faster than ItISC and Huffman on both sides. Overall, these results suggest that VCSC is suitable for smart-city deployments with moderately capable edge devices, while placing the heavier reconstruction workload on the data center receiver.

D. Ablation Study

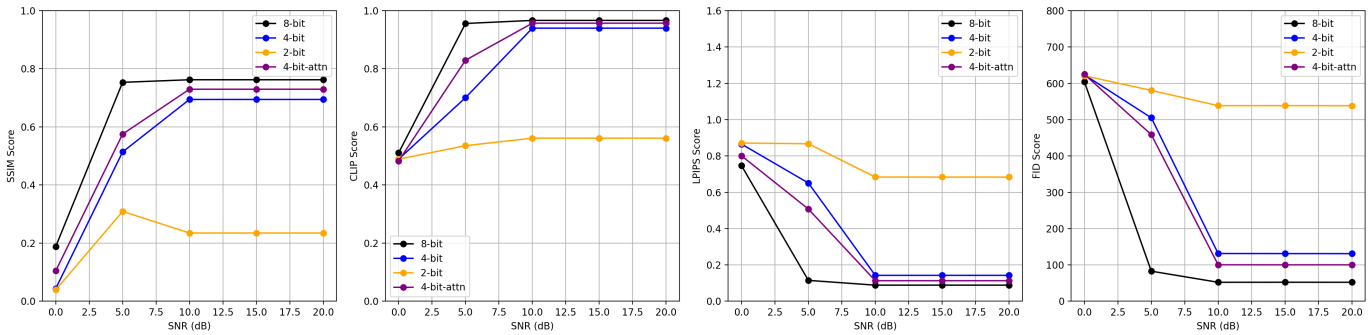
To further investigate the performance of the proposed VCSC, an ablation study was conducted using the SmartCity Dataset. This study systematically analyses key components of the method, including the impact of different quantization levels, the role of Turbo coding, and the resulting visual quality of reconstructed images.

1) Quantization: Fig. 6 presents an ablation study examining the impact of three uniform quantization levels (8-bit, 4-bit, and 2-bit) alongside a single attention-based variant (4-bit-attn) with a threshold of 0.2 under AWGN (Fig. 6a) and Rayleigh (Fig. 6b) channels.

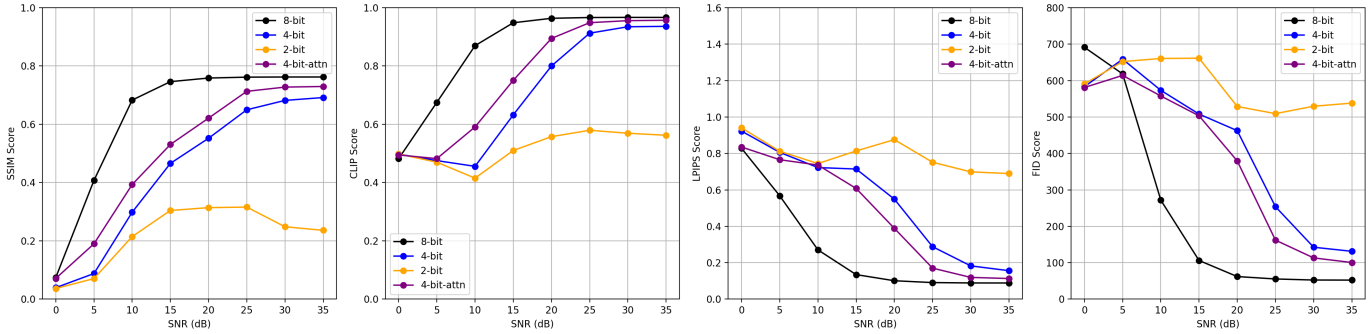
Under the AWGN channel, VCSC (8-bit) achieves high SSIM and CLIP scores and low LPIPS and FID scores, indicating its effectiveness in preserving both structural and semantic information. VCSC (4-bit) and VCSC (2-bit) maintain stable performance at moderate to high SNR values while reducing the total bit rate. VCSC (4-bit-attn) demonstrates effective noise mitigation by allocating more bits to high-attention regions, leading to improved robustness in challenging conditions.

In the Rayleigh fading channel scenario, VCSC (4-bit-attn) maintains stability compared to VCSC (4-bit). While VCSC (2-bit) faces challenges under fading conditions, VCSC (8-bit) provides consistent reconstruction quality across all SNR levels. These results suggest that leveraging attention maps for adaptive quantization contributes to a balance between

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549

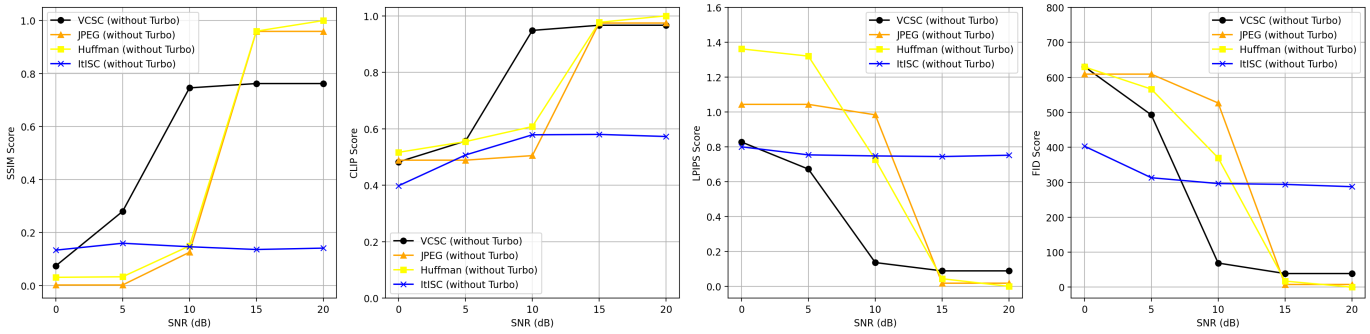


(a) AWGN channel.

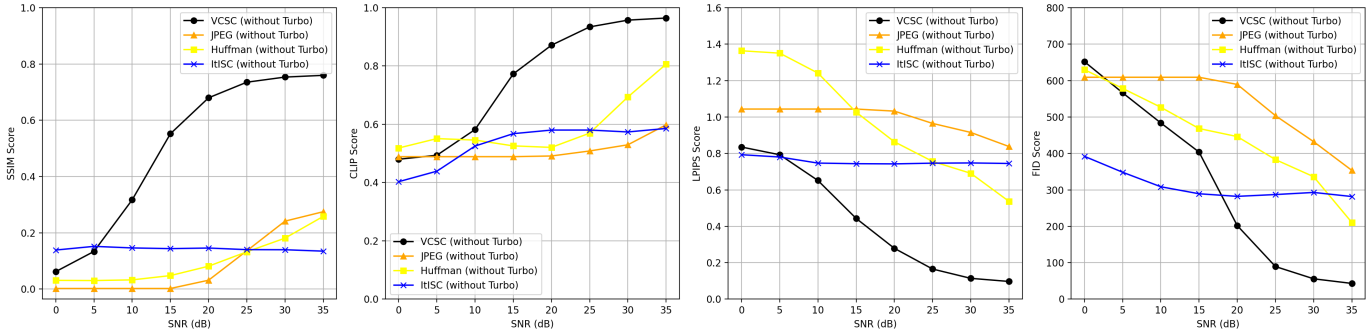


(b) Rayleigh fading channel.

Fig. 6: Impact of uniform quantization (8-bit, 4-bit, 2-bit) and 4-bit attention-based quantization (threshold 0.2) on VCSC performance under (a) AWGN and (b) Rayleigh fading channels.



(a) AWGN channel.



(b) Rayleigh fading channel.

Fig. 7: Performance comparison of VCSC and baselines (JPEG, Huffman, ItISC) without Turbo coding under (a) AWGN and (b) Rayleigh fading channels.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549

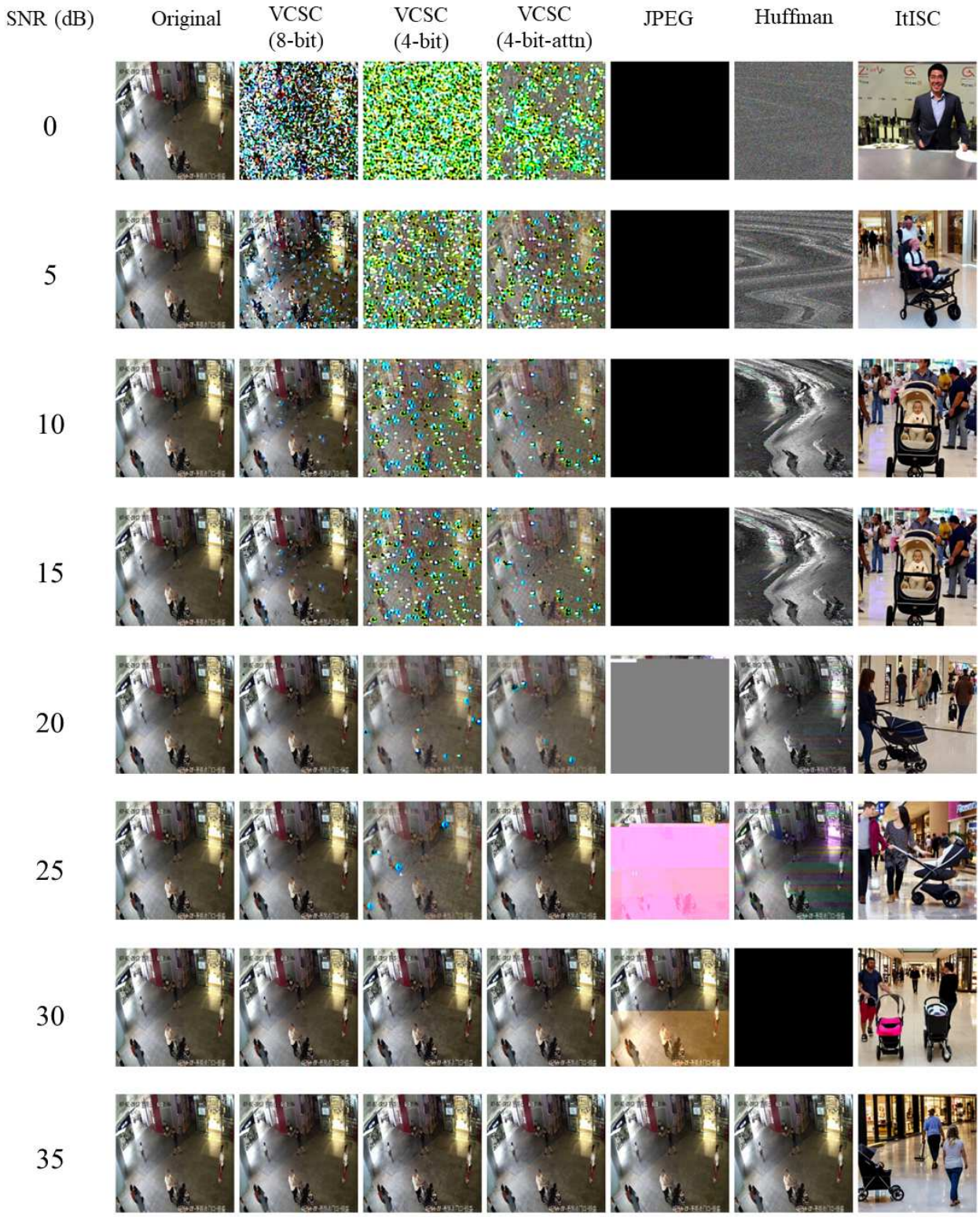


Fig. 8: Visual comparison of reconstructed images from VCSC variants (8-bit, 4-bit, 4-bit-attn) and baseline methods (JPEG, Huffman, ItISC) across SNR levels from 0 dB to 35 dB on the SmartCity Dataset.

compression efficiency and reconstruction quality in wireless transmission.

Table IV summarizes the compression rates for different quantization methods on the SmartCity Dataset. As expected, lower-bit quantization results in greater compression, with VCSC (2-bit) providing the most reduction (0.5%) but at the cost of performance degradation. The VCSC (4-bit-attn) method strikes a balance between compression and reconstruction quality: although its compression rate (1.7%) is higher (i.e., less compression) than standard VCSC (4-bit, 1.0%), it delivers better reconstruction due to adaptive bit allocation. By contrast, VCSC (8-bit) achieves a compression rate of 2.1%, ensuring strong preservation of structural and semantic information. These findings highlight the trade-off between compression and reconstruction quality, reinforcing the advantages of attention-based quantization in challenging scenarios.

TABLE IV: Compression rates of VCSC with varying quantization schemes, where the compression rate is computed as $\frac{\text{Compressed Bits}}{\text{Original Bits}}$. A smaller value means stronger compression.

Method	Compression Rate
VCSC (8-bit)	2.1%
VCSC (4-bit-attn)	1.7%
VCSC (4-bit)	1.0%
VCSC (2-bit)	0.5%

2) Without Turbo: Fig. 7 illustrates the performance of four methods, VCSC (8-bit), JPEG, Huffman, and ItISC, when Turbo coding is removed, under both AWGN (Fig. 7a) and Rayleigh (Fig. 7b) channels. SSIM, CLIP Score, LPIPS, and FID are reported across varying SNR values.

In the AWGN channel, VCSC maintains stable SSIM, CLIP, LPIPS, and FID scores, indicating that latent-space transmission provides strong resilience against noise. JPEG, Huffman, and ItISC exhibit rapid performance degradation, with Huffman being particularly sensitive to channel impairments.

Under the Rayleigh fading channel, VCSC continues to demonstrate resilience across SSIM, CLIP, LPIPS, and FID scores, maintaining relatively stable performance in challenging conditions. These findings highlight the effectiveness of VCSC's latent-space approach in scenarios where Turbo coding is absent.

3) Visualization: Fig. 8 provides a visual comparison of reconstructed images produced by various methods across a range of SNR values (0 dB to 35 dB). The figure includes results from VCSC (8-bit), VCSC (4-bit), and VCSC (4-bit-attn) with a threshold of 0.2, alongside baseline methods such as JPEG, Huffman, and ItISC. Several key observations can be drawn from the visualization:

- VCSC (8-bit). VCSC (8-bit) effectively preserves structural details and maintains clear semantic content across different SNR values. The reconstructed images exhibit minimal artifacts and maintain a strong resemblance to the original input.
- VCSC (4-bit) vs. VCSC (4-bit-attn). While VCSC (4-bit) shows some degradation due to reduced bit depth, VCSC (4-bit-attn) mitigates these effects by allocating higher precision to regions of importance. As a result, VCSC (4-bit-attn) retains structural clarity and semantic information, particularly in challenging channel conditions.
- Baseline Methods. JPEG, Huffman, and ItISC experience significant quality degradation at low SNR levels. Artifacts become noticeable, and fine details are lost, indicating low resilience to channel noise. In contrast, VCSC methods demonstrate smooth transitions and reduced distortions as SNR increases.
- SNR Impact. At low SNR values (e.g., 0 dB to 10 dB), all methods exhibit degradation; however, VCSC methods, particularly VCSC (8-bit) and VCSC (4-bit-attn), retain recognizable features. As SNR improves (above 25 dB), the differences in reconstruction quality diminish, but VCSC methods continue to produce images that are perceptually close to the original compared to the baselines.

Overall, the visual comparison highlights the effectiveness of the proposed VCSC, particularly when enhanced by attention-based quantization, in preserving both structural and semantic information under noisy channel conditions.

VI. Conclusion

This paper introduced the VCSC system tailored for efficient image transmission in smart city environments. Unlike conventional compression and text-centered semantic communication methods, VCSC directly encodes images into a compact latent code using a pretrained encoder–decoder network, ensuring high-fidelity reconstructions while significantly reducing bandwidth requirements. The proposed VCSC also integrates an attention-based quantization strategy to dynamically allocate bit precision to semantically important regions, enhancing compression efficiency without sacrificing perceptual quality. Comprehensive evaluations across AWGN and Rayleigh fading channel models demonstrate that VCSC consistently outperforms traditional methods such as JPEG, Huffman coding, and text-centered image transmission in terms of SSIM, CLIP, LPIPS, and FID scores. The results highlight VCSC's robustness against channel impairments and its ability to preserve both structural and semantic information, making it well-suited for real-world applications such as smart city surveillance, traffic monitoring, and infrastructure management.

Future work will focus on refining the proposed quantization strategy to further improve the trade-off between compression efficiency and reconstruction fidelity, particularly under dynamic wireless channel conditions. In addition, we will investigate adaptive coding and bit-allocation mechanisms to enhance transmission robustness and efficiency across varying SNR regimes.

References

[1] W. Ji, J. Xu, H. Qiao, M. Zhou, and B. Liang, "Visual IoT: Enabling internet of things visualization in smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 102–110, 2019.

[2] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[3] S. Dhawan, "A review of image compression and comparison of its algorithms," *International Journal of Electronics & Communication Technology*, vol. 2, no. 1, pp. 22–26, 2011.

[4] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4565–4574.

[5] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, 2021.

[6] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 2022.

[7] J. Dai, P. Zhang, K. Niu, S. Wang, Z. Si, and X. Qin, "Communication beyond transmitting bits: Semantics-guided source and channel coding," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 170–177, 2022.

[8] S. Liu, Z. Peng, Q. Yu, and L. Duan, "A novel image semantic communication method via dynamic decision generation network and generative adversarial network," *Scientific Reports*, vol. 14, no. 1, p. 19636, 2024.

[9] T. Han, J. Tang, Q. Yang, Y. Duan, Z. Zhang, and Z. Shi, "Generative model based highly efficient semantic communication approach for image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* IEEE, 2023, pp. 1–5.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[11] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2645–2657, 2023.

[12] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019.

[13] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, 2021.

[14] H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, "Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* IEEE, 2024, pp. 13 506–13 510.

[15] F. Jiang, C. Tang, L. Dong, K. Wang, K. Yang, and C. Pan, "Visual language model based cross-modal semantic communication systems," *IEEE Trans. Wireless Commun.*, 2025.

[16] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5306–5314.

[17] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021.

[19] Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Generative latent coding for ultra-low bitrate image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26 088–26 098.

[20] Q. Fu, H. Xie, Z. Qin, G. Slabaugh, and D. Tao, "Vector quantized semantic communication system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 982–986, 2023.

[21] R. M. Gray and D. L. Neuhoff, *Quantization*. IEEE Press, 1998.

[22] J. Ballé, D. Minnen, S. J. Singh, S. Hwang, and N. Johnston, "Variational image compression with a hyperprior," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[23] Z. Zhong, H. Akutsu, and K. Aizawa, "Channel-level variable quantization network for deep image compression," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 465–471.

[24] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 873–12 883.

[25] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, "Intelligent edge computing for IoT-based energy management in smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 111–117, 2019.

[26] X. Li, M. Zhao, M. Zeng, S. Mumtaz, V. G. Menon, Z. Ding, and O. A. Dobre, "Hardware impaired ambient backscatter NOMA systems: Reliability and security," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2723–2736, 2021.

[27] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, 1996.

[28] W. Tan, Y. Wang, L. Liu, X. Wang, and T. Ding, "Adaptive federated deep learning-based semantic communication in the social internet of things," *IEEE Internet Things J.*, 2024.

[29] X. Li, Y. Zheng, W. U. Khan, M. Zeng, D. Li, G. K. Ragesh, and L. Li, "Physical layer security of cognitive ambient backscatter communications for green internet-of-things," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1066–1076, 2021.

[30] J. Wang, S. Wang, J. Dai, Z. Si, D. Zhou, and K. Niu, "Perceptual learned source-channel coding for high-fidelity image semantic transmission," in *Proc. IEEE Global Commun. Conf.* IEEE, 2022, pp. 3959–3964.

[31] K. Guo, Y. Lu, H. Gao, and R. Cao, "Artificial intelligence-based semantic internet of things in a user-centric smart city," *Sensors*, vol. 18, no. 5, p. 1341, 2018.

[32] S. Pandya, G. Srivastava, R. Jhaveri, M. R. Babu, S. Bhattacharya, P. K. R. Maddikunta, S. Mastorakis, M. J. Piran, and T. R. Gadekallu, "Federated learning for smart cities: A comprehensive survey," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102987, 2023.

[33] X. Li, Q. Wang, M. Zeng, Y. Liu, S. Dang, T. A. Tsiftsis, and O. A. Dobre, "Physical-layer authentication for ambient backscatter-aided NOMA symbiotic systems," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2288–2303, 2023.

[34] A. Razi, X. Chen, H. Li, H. Wang, B. Russo, Y. Chen, and H. Yu, "Deep learning serves traffic safety analysis: A forward-looking review," *IET Intelligent Transport Systems*, vol. 17, no. 1, pp. 22–71, 2023.

[35] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, 2016.

[36] X. Chen, C. Huang, G. Chen, D. Feng, and P. Xiao, "The communication and computation trade-off in wireless semantic communications," *IEEE Wireless Commun. Lett.*, 2025.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[41] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2022, pp. 12 888–12 900.

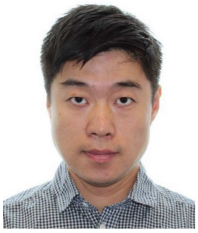
[42] H. Xie, Z. Qin, G.-Y. Li, and B.-S. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[43] Y. B. Sardoğan, "Traffic detection project," 2023, accessed: 2025-03-04. [Online]. Available: https://www.kaggle.com/datasets/yusufberksardoan/traffic-detection-project

[44] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.* IEEE, 2018, pp. 1113–1121.

[45] D. Pudasaini and A. Abhari, "Edge-based video analytic for smart cities," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2026.3656549



Yan Gong is a Lecturer in Computer Science at the School of Computing and Engineering, Bournemouth University, UK. He received his PhD in Computer Science from Loughborough University in 2023. His primary research interests include NLP, cross-modal learning, generative AI, and AI agents. Prior to his PhD, he gained over six years of industry experience as an AI engineer, developing a strong understanding of practical AI applications and their impact on the technology sector.

Zheng Chu (Member, IEEE) is an Assistant Professor with Department of Electrical and Electronic Engineering, The University of Nottingham Ningbo China. Prior to this, he held positions in University of Surrey, from 2017 to 2024, and Middlesex University, from 2016 to 2017, respectively. He received MSc and Ph.D. degrees from Newcastle University, Newcastle-upon Tyne, UK, in 2012 and 2016, respectively. His current research interests include 6G communications, IoT networks, Rydberg atomic receiver, polarforming antenna, Space-air-ground integrated networks, smart mobility and transportation, as well as AI-empowered future networks. He received the Exemplary Reviewer for IEEE TRANSACTIONS ON COMMUNICATIONS in 2022, and the Best Paper Awards of IEEE/CIC UCOM (2024 and 2025), IEEE ICCT (2024), and EAI CHINACOM (2024).

Zhengyu Zhu (Senior Member, IEEE) received the Ph.D. degree in information engineering from Zhengzhou University, Zhengzhou, China, in 2017. From October 2013 to October 2015, he visited the Communication and Intelligent System Laboratory, Korea University, Seoul, South Korea, to conduct a collaborative research as a Visiting Student. He is currently a professor with Zhengzhou University. He served as an Associate Editor for the IEEE SENSOR JOURNAL, IEEE SYSTEMS JOURNAL, JOURNAL OF COMMUNICATIONS AND NETWORKS, the PHYSICAL COMMUNICATIONS. His research interests include information theory and signal processing for wireless communications such as B5G/6G, Intelligent reflecting surface, Internet of Things, machine learning, millimeter wave communication, UAV communication, physical layer security, convex optimization techniques, and energy harvesting communication systems.

Pei Xiao (Senior Member, IEEE) is currently a Professor of wireless communications with the Institute for Communication Systems, University of Surrey, Guildford, U.K. He is also a Technical Manager of 5GIC/6GIC, leading the research team in the new physical layer work area, and coordinating/supervising research activities across all the work areas (https://www.surrey.ac.uk/institute-communicationsystems/5g-6g-innovation-centre). He was with Newcastle University and Queen's University Belfast. He also held positions with Nokia Networks, Finland. He has authored or coauthored extensively in the fields of communication theory, RF and antenna design, signal processing for wireless communications, and is an inventor on more than 15 recent patents addressing bottleneck problems in 5G/6G systems.

Ming Zeng (Member, IEEE) received the B.E. and master's degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and 2016, respectively, and the Ph.D. degree in telecommunications engineering from the Memorial University of Newfoundland, St. John's, NL, Canada, in 2020. He is currently an Associate Professor and a Canada Research Chair with the Department of Electrical and Computer Engineering, Laval University, Quebec City, QC, Canada. He has published more than 140 articles and conferences in first-tier IEEE journals and proceedings, and his work has been cited over 5900 times per Google Scholar. His research interests include resource allocation for beyond 5G systems and machine learning-empowered optical communications. He serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and IEEE WIRELESS COMMUNICATIONS LETTERS.

Yi Wang is currently an Associate Professor with the School of Electronics and Information, Zhengzhou University of Aeronautics, China. He received B.S. degree from Information Engineering University, Zhengzhou, China, in 2006, and the M.S. and Ph.D. degrees from the school of information science and engineering, Southeast University, China, in 2009 and 2016, respectively. His current research interests include massive MIMO, energy-efficient communication, UAV-aided communication, physical layer security, wireless power transfer and intelligent reflecting surface-aided wireless communication. He received the best paper awards of the IEEE WCSP in 2015.

Hari Pandey is a Senior Lecturer in Data Science and Artificial Intelligence at the School of Computing and Engineering, Bournemouth University, UK. He is featured in Stanford University's 2021-2025 list of the world's top 2% scientists. Hari's expertise lies in Computer Science & Engineering, with research interests in artificial intelligence, machine learning, deep learning, natural language processing, Large Language Models (LLMs), soft computing, and computer vision. He has authored several books, including State of the Art on Grammatical Inference Using Evolutionary Method (Elsevier), and has published over 180 papers in leading journals and conferences. He serves on the editorial boards of major journals such as Neural Networks, Applied Soft Computing, and multiple IEEE Transactions, and regularly reviews for top international conferences. He is a Fellow of the UK Higher Education Academy, an award-winning academic, and has delivered keynote and invited talks worldwide. Previously, he held academic and research positions at Edge Hill University and Middlesex University London, where he worked as a Research Fellow on the EU-funded DREAM4CARS project.

Jingrui Hou received his PhD from Loughborough University, UK. He is currently a Postdoctoral Researcher at the School of Information Management, Wuhan University. His primary research interests include artificial intelligence governance and natural language processing.